group13model

group13

2023-03-16

Data Loading and Pre-processing

Removing Value

We noticed missing data and outliers in the data cleaning section, so we removed them before building our model.

```
dataset13 <- read_csv("dataset13.csv")
newdataset<- na.omit(dataset13)
newdataset<- newdataset%>%
  arrange(desc(altitude_mean_meters))
newdataset<- newdataset[-c(1:4),]
newdataset<- newdataset%>%
  arrange(aroma)
newdataset<- newdataset[-1,]
str(newdataset)
```

Calculating Correlation

In order to prepare for subsequent improvement and selection of variables during modelling, we firstly calculated the correlation between every two numerical variables.

```
newdataset[,2:6]%>%
  cor()%>%
  kable(caption='\\label{tab:correlation} correlation between 5 numerical variables')%>%
  kable_styling(font_size = 10, latex_options = "hold_position")
```

Table 1: correlation between 5 numerical variables

	aroma	flavor	acidity	category_two_defects	altitude_mean_meters
aroma	1.0000000	0.7253135	0.5907547	-0.1934092	0.1632542
flavor	0.7253135	1.0000000	0.7438336	-0.2477485	0.1476604
acidity	0.5907547	0.7438336	1.0000000	-0.1851076	0.1778057
category_two_defects	-0.1934092	-0.2477485	-0.1851076	1.0000000	-0.0025717
altitude_mean_meters	0.1632542	0.1476604	0.1778057	-0.0025717	1.0000000

Table 1 shows the correlation between every two variables including aroma,flavor,acidity,category_two_defects and altitude_mean_meters.We can see that the correlation between aroma and flavor (0.72) and the correlation between flavor acidity (0.74) are both more than 0.7, which means they are of strong positive

correlation, there is also a moderate correlation between aroma and acidity (0.59), while the correlation between other pairs are relatively weak.

Processing Non-numerical Data

For non-numerical data, including country_of_origin, Qualityclass and harvested, we set the country_of_origin and harvested as factors, while as a qualitative variable, we converted Qualityclass into dummy variables, 'poor' to '0' and 'good' to '1'.

```
names(newdataset)
newdataset$country_of_origin<- as.factor(newdataset$country_of_origin)
newdataset$Qualityclass<- ifelse(newdataset$Qualityclass=='Poor',0,1)
newdataset$harvested <- as.factor(newdataset$harvested)
```

Formal Data Analysis

We used GLM to fit a logistic regression model with Qualityclass as the binary response variable, and country_of_origin, aroma, flavor, acidity, category_two_defects, altitude_mean_meaters and harvested as the explanatory variables. A summary of the model and the a graph showing the points estimate for the log-odds with their corresponding 95% confidence interval are obtained as results.

Basic GLM

```
mod.cafe <- glm(Qualityclass ~  country_of_origin +aroma + flavor+acidity+category_two_defects+altitude
                family = binomial(link = "logit"))
summary(mod.cafe) #AIC 543
```

```
Call:
glm(formula = Qualityclass ~ country_of_origin + aroma + flavor +
    acidity + category_two_defects + altitude_mean_meters + harvested,
    family = binomial(link = "logit"), data = newdataset)

Deviance Residuals:
    Min       1Q    Median       3Q       Max
-4.5339   -0.2406    0.0000    0.2851    3.6011

Coefficients:
                                     Estimate Std. Error z value
(Intercept)                        -1.446e+02  1.164e+01 -12.416
country_of_originBurundi            1.908e+00  4.926e+00   0.387
country_of_originChina              4.996e-01  1.081e+00   0.462
country_of_originColombia           1.816e+00  5.642e-01   3.218
country_of_originCosta Rica         2.898e-01  7.635e-01   0.380
country_of_originCote d?Ivoire     -1.211e+01  6.523e+03  -0.002
country_of_originEcuador           -1.427e+00  1.495e+00  -0.954
country_of_originEl Salvador        5.411e-01  9.579e-01   0.565
country_of_originEthiopia           1.333e+01  8.981e+02   0.015
country_of_originGuatemala         -5.834e-01  5.456e-01  -1.069
```

```
country_of_originHaiti                              2.125e+00  1.788e+00   1.189
country_of_originHonduras                          -3.959e-01  6.791e-01  -0.583
country_of_originIndia                             -2.988e+00  1.085e+00  -2.754
country_of_originIndonesia                         -1.188e-01  9.863e-01  -0.120
country_of_originKenya                              6.488e-01  1.596e+00   0.406
country_of_originLaos                              -1.513e+01  4.504e+03  -0.003
country_of_originMalawi                            -8.307e-01  1.284e+00  -0.647
country_of_originMauritius                         -1.206e+01  6.523e+03  -0.002
country_of_originMexico                            -6.986e-01  5.074e-01  -1.377
country_of_originMyanmar                           -1.634e+01  2.361e+03  -0.007
country_of_originNicaragua                          5.066e-01  1.977e+00   0.256
country_of_originPanama                             3.344e+00  1.820e+00   1.837
country_of_originPeru                              -1.416e+01  6.523e+03  -0.002
country_of_originPhilippines                        2.892e+00  2.778e+00   1.041
country_of_originTaiwan                             7.848e-01  6.631e-01   1.184
country_of_originTanzania, United Republic Of       1.023e+00  7.536e-01   1.358
country_of_originThailand                           2.560e+00  9.800e-01   2.612
country_of_originUganda                            -1.550e+00  7.903e-01  -1.962
country_of_originUnited States                     -3.077e-01  1.678e+00  -0.183
country_of_originUnited States (Hawaii)             4.261e+00  6.523e+03   0.001
country_of_originUnited States (Puerto Rico)       -3.300e+00  1.711e+00  -1.929
country_of_originVietnam                            2.168e+00  1.162e+00   1.865
country_of_originZambia                            -1.363e+01  6.523e+03  -0.002
aroma                                               5.188e+00  8.479e-01   6.119
flavor                                              8.556e+00  1.062e+00   8.060
acidity                                             5.230e+00  8.242e-01   6.346
category_two_defects                                5.265e-02  3.506e-02   1.502
altitude_mean_meters                                5.736e-04  3.216e-04   1.784
harvested2011                                      -2.323e-01  1.122e+00  -0.207
harvested2012                                       7.648e-02  9.692e-01   0.079
harvested2013                                       6.018e-01  9.772e-01   0.616
harvested2014                                       4.134e-02  9.858e-01   0.042
harvested2015                                      -6.614e-02  9.757e-01  -0.068
harvested2016                                       7.450e-01  1.022e+00   0.729
harvested2017                                       5.002e-01  1.028e+00   0.487
harvested2018                                       2.027e+00  1.264e+00   1.604
                                                   Pr(>|z|)
(Intercept)                                         < 2e-16 ***
country_of_originBurundi                            0.69845
country_of_originChina                              0.64390
country_of_originColombia                           0.00129 **
country_of_originCosta Rica                         0.70422
country_of_originCote d?Ivoire                      0.99852
country_of_originEcuador                            0.33999
country_of_originEl Salvador                        0.57214
country_of_originEthiopia                           0.98816
country_of_originGuatemala                          0.28488
country_of_originHaiti                              0.23453
country_of_originHonduras                           0.55989
country_of_originIndia                              0.00589 **
country_of_originIndonesia                          0.90415
country_of_originKenya                              0.68442
country_of_originLaos                               0.99732
country_of_originMalawi                             0.51769
```

```
country_of_originMauritius                       0.99852
country_of_originMexico                          0.16855
country_of_originMyanmar                         0.99448
country_of_originNicaragua                       0.79777
country_of_originPanama                          0.06615 .
country_of_originPeru                            0.99827
country_of_originPhilippines                     0.29788
country_of_originTaiwan                          0.23660
country_of_originTanzania, United Republic Of    0.17453
country_of_originThailand                        0.00899 **
country_of_originUganda                          0.04979 *
country_of_originUnited States                   0.85452
country_of_originUnited States (Hawaii)          0.99948
country_of_originUnited States (Puerto Rico)     0.05376 .
country_of_originVietnam                         0.06212 .
country_of_originZambia                          0.99833
aroma                                            9.43e-10 ***
flavor                                           7.64e-16 ***
acidity                                          2.21e-10 ***
category_two_defects                             0.13321
altitude_mean_meters                             0.07449 .
harvested2011                                    0.83603
harvested2012                                    0.93711
harvested2013                                    0.53798
harvested2014                                    0.96655
harvested2015                                    0.94595
harvested2016                                    0.46625
harvested2017                                    0.62654
harvested2018                                    0.10876
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1289.15  on 929  degrees of freedom
Residual deviance:  451.69  on 884  degrees of freedom
AIC: 543.69

Number of Fisher Scoring iterations: 17
```

```
plot_model(mod.cafe, show.values = TRUE,
           title = "", show.p = FALSE, value.offset = 0.5)
```
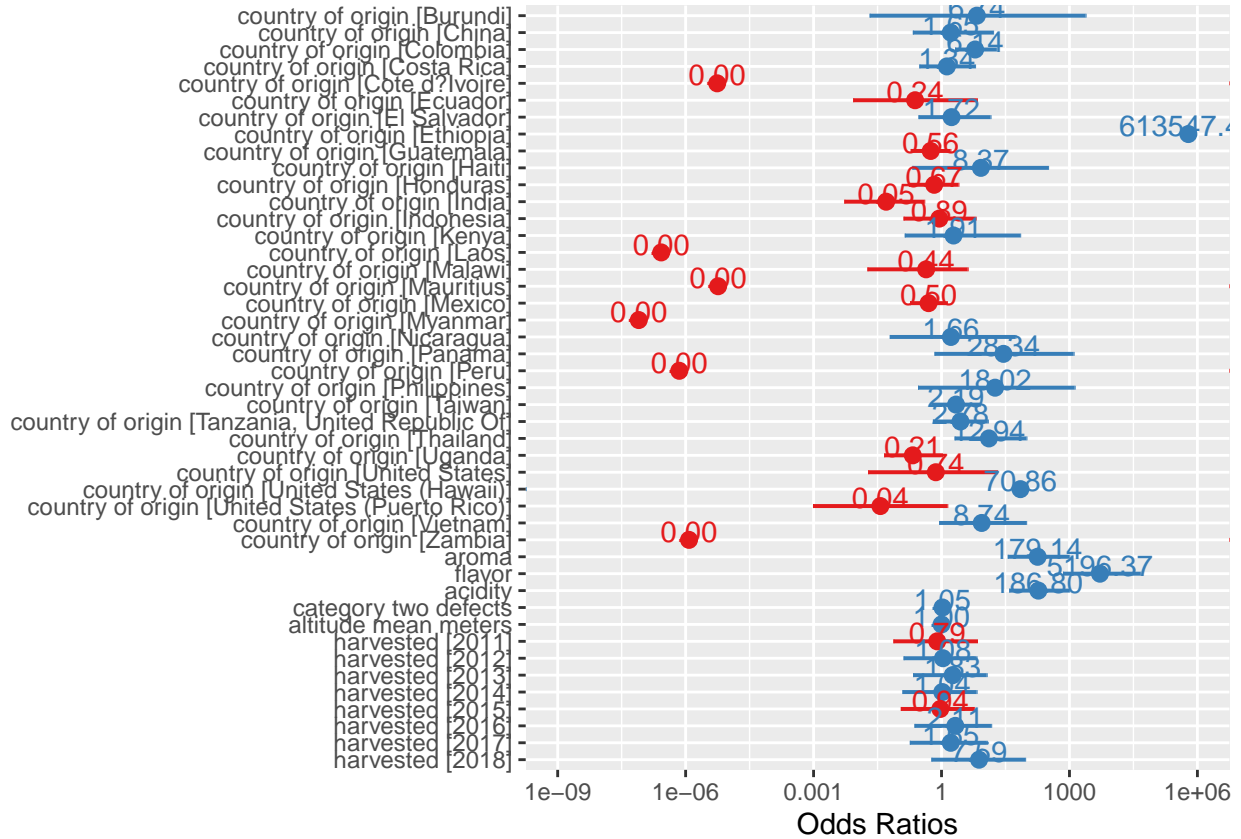

Figure 1: Odds of various factors influencing the quality of coffee(basic GLM)

In the results we can see that aroma, flavor and acidity has coefficients of 5.19, 8.56, 5,23 separately, indicating comparatively strong positive influence on cafe quality, whilst category_two_defects and altitude_mean_meters do not appear to have much impact. For country_of_origin and harvested, different countries and vintages have different degrees of influence on the quality of coffee. For example, Ethiopia has a strong capacity to produce good coffee, as it has a coefficient more than ten(13.33) , Panama and Hawaii have lower coefficients(3.34 and 4.26), but still can be a good places to make coffee. However, there are countries with coefficients below -10,like Cote d'Ivoire,Laos,Mauritius, Myanmar,Peru and Zambia, which shows that they are likely to produce poorer cafe. In addition, only the harvested of 2018 shows a little positive impact on cafe quality(2.03), while other variables do not appear to be strongly influential.

GLM Stepwise

In the previous basic GLM we fitted a model with AIC of 543, wondering whether there is better regression to fit the data after selecting only the influencial variables, we then decided to use stepwise regression to improve our model.

```
# Fit a glm using stepwise regression with AIC as the criterion
model.step <- stepAIC(glm(Qualityclass ~ country_of_origin +aroma + flavor+acidity+category_two_defect
summary(mod.cafe) #AIC537
```

```
Call:
glm(formula = Qualityclass ~ country_of_origin + aroma + flavor +
    acidity + category_two_defects + altitude_mean_meters + harvested,
    family = binomial(link = "logit"), data = newdataset)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.5339  -0.2406   0.0000   0.2851   3.6011

Coefficients:
                                                Estimate Std. Error z value
(Intercept)                                   -1.446e+02  1.164e+01 -12.416
country_of_originBurundi                       1.908e+00  4.926e+00   0.387
country_of_originChina                         4.996e-01  1.081e+00   0.462
country_of_originColombia                      1.816e+00  5.642e-01   3.218
country_of_originCosta Rica                    2.898e-01  7.635e-01   0.380
country_of_originCote d?Ivoire                -1.211e+01  6.523e+03  -0.002
country_of_originEcuador                      -1.427e+00  1.495e+00  -0.954
country_of_originEl Salvador                   5.411e-01  9.579e-01   0.565
country_of_originEthiopia                      1.333e+01  8.981e+02   0.015
country_of_originGuatemala                    -5.834e-01  5.456e-01  -1.069
country_of_originHaiti                         2.125e+00  1.788e+00   1.189
country_of_originHonduras                     -3.959e-01  6.791e-01  -0.583
country_of_originIndia                        -2.988e+00  1.085e+00  -2.754
country_of_originIndonesia                    -1.188e-01  9.863e-01  -0.120
country_of_originKenya                         6.488e-01  1.596e+00   0.406
country_of_originLaos                         -1.513e+01  4.504e+03  -0.003
country_of_originMalawi                       -8.307e-01  1.284e+00  -0.647
country_of_originMauritius                    -1.206e+01  6.523e+03  -0.002
country_of_originMexico                       -6.986e-01  5.074e-01  -1.377
country_of_originMyanmar                      -1.634e+01  2.361e+03  -0.007
country_of_originNicaragua                     5.066e-01  1.977e+00   0.256
country_of_originPanama                        3.344e+00  1.820e+00   1.837
country_of_originPeru                         -1.416e+01  6.523e+03  -0.002
country_of_originPhilippines                   2.892e+00  2.778e+00   1.041
country_of_originTaiwan                        7.848e-01  6.631e-01   1.184
country_of_originTanzania, United Republic Of  1.023e+00  7.536e-01   1.358
country_of_originThailand                      2.560e+00  9.800e-01   2.612
country_of_originUganda                       -1.550e+00  7.903e-01  -1.962
country_of_originUnited States                -3.077e-01  1.678e+00  -0.183
country_of_originUnited States (Hawaii)        4.261e+00  6.523e+03   0.001
country_of_originUnited States (Puerto Rico)  -3.300e+00  1.711e+00  -1.929
country_of_originVietnam                       2.168e+00  1.162e+00   1.865
```

```
country_of_originZambia                                  -1.363e+01  6.523e+03   -0.002
aroma                                                     5.188e+00  8.479e-01    6.119
flavor                                                    8.556e+00  1.062e+00    8.060
acidity                                                   5.230e+00  8.242e-01    6.346
category_two_defects                                      5.265e-02  3.506e-02    1.502
altitude_mean_meters                                      5.736e-04  3.216e-04    1.784
harvested2011                                            -2.323e-01  1.122e+00   -0.207
harvested2012                                             7.648e-02  9.692e-01    0.079
harvested2013                                             6.018e-01  9.772e-01    0.616
harvested2014                                             4.134e-02  9.858e-01    0.042
harvested2015                                            -6.614e-02  9.757e-01   -0.068
harvested2016                                             7.450e-01  1.022e+00    0.729
harvested2017                                             5.002e-01  1.028e+00    0.487
harvested2018                                             2.027e+00  1.264e+00    1.604
                                                          Pr(>|z|)
(Intercept)                                               < 2e-16 ***
country_of_originBurundi                                  0.69845
country_of_originChina                                    0.64390
country_of_originColombia                                 0.00129 **
country_of_originCosta Rica                               0.70422
country_of_originCote d?Ivoire                            0.99852
country_of_originEcuador                                  0.33999
country_of_originEl Salvador                              0.57214
country_of_originEthiopia                                 0.98816
country_of_originGuatemala                                0.28488
country_of_originHaiti                                    0.23453
country_of_originHonduras                                 0.55989
country_of_originIndia                                    0.00589 **
country_of_originIndonesia                                0.90415
country_of_originKenya                                    0.68442
country_of_originLaos                                     0.99732
country_of_originMalawi                                   0.51769
country_of_originMauritius                                0.99852
country_of_originMexico                                   0.16855
country_of_originMyanmar                                  0.99448
country_of_originNicaragua                                0.79777
country_of_originPanama                                   0.06615 .
country_of_originPeru                                     0.99827
country_of_originPhilippines                              0.29788
country_of_originTaiwan                                   0.23660
country_of_originTanzania, United Republic Of             0.17453
country_of_originThailand                                 0.00899 **
country_of_originUganda                                   0.04979 *
country_of_originUnited States                            0.85452
country_of_originUnited States (Hawaii)                   0.99948
country_of_originUnited States (Puerto Rico)              0.05376 .
country_of_originVietnam                                  0.06212 .
country_of_originZambia                                   0.99833
aroma                                                     9.43e-10 ***
flavor                                                    7.64e-16 ***
acidity                                                   2.21e-10 ***
category_two_defects                                      0.13321
altitude_mean_meters                                      0.07449 .
harvested2011                                             0.83603
```

```
harvested2012                                0.93711
harvested2013                                0.53798
harvested2014                                0.96655
harvested2015                                0.94595
harvested2016                                0.46625
harvested2017                                0.62654
harvested2018                                0.10876
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1289.15  on 929  degrees of freedom
Residual deviance:  451.69  on 884  degrees of freedom
AIC: 543.69

Number of Fisher Scoring iterations: 17
```

```
plot_model(model.step, show.values = TRUE,
           title = "", show.p = FALSE, value.offset = 0.50)
```
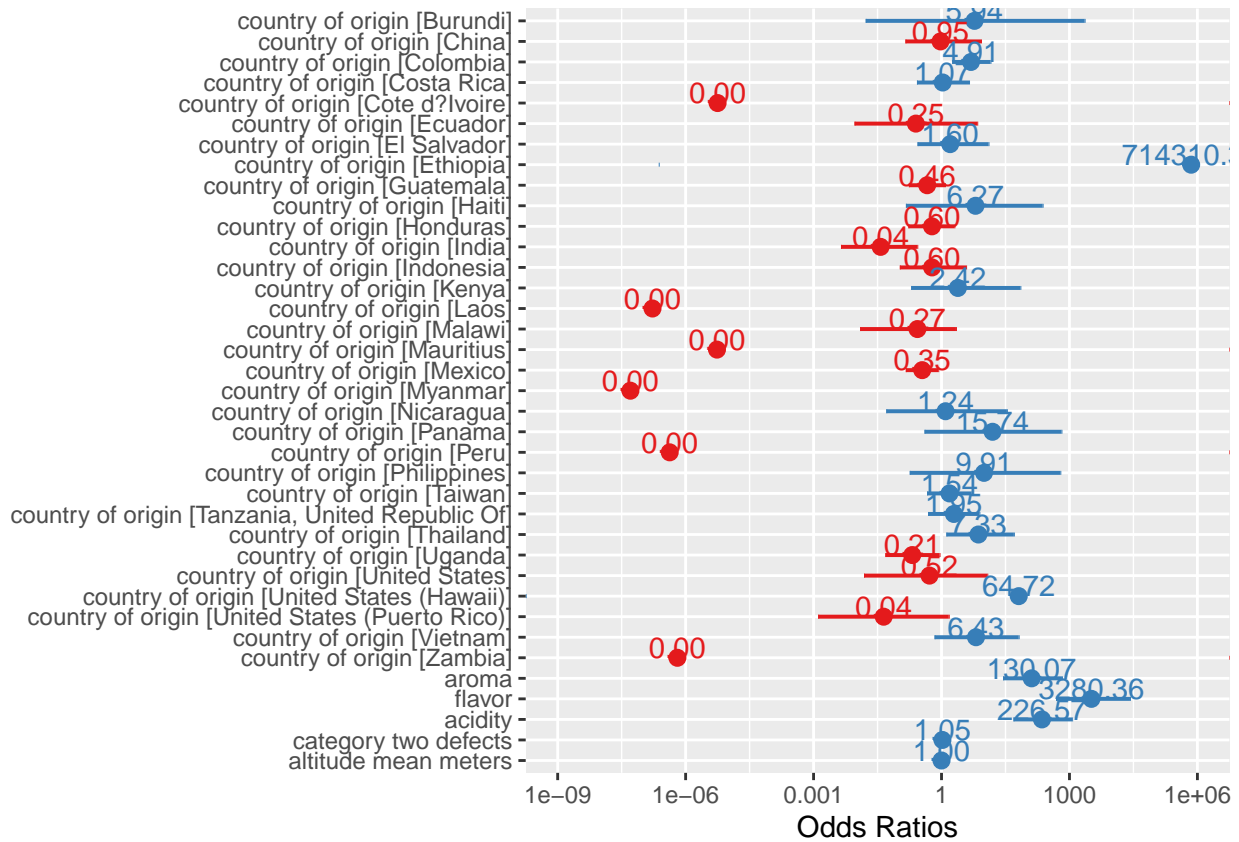

Figure 2: Odds of various factors influencing the quality of coffee(GLM stepwise regression)

Using stepwise regression, we fitted a model with AIC of 537, which is relatively smaller than 543 in our first basic model, hence we can say that stepwise regression helped us to improve our model.

Adding Interaction Terms

Considering the possible interactions between the variables, based on the previously calculated correlations,we added some interaction terms(aroma:flavor,flavor:acidity,aroma:acidity)in order to improve our model.We summarized the results and graphically showed the log-odds and their corresponding 95% confidence intervals.

```
mod.cafe <- glm(Qualityclass ~  country_of_origin +aroma + flavor+acidity+category_two_defects+altitude_
                family = binomial(link = "logit"))
summary(mod.cafe) #AIC539
```

```
Call:
glm(formula = Qualityclass ~ country_of_origin + aroma + flavor +
    acidity + category_two_defects + altitude_mean_meters + harvested +
    aroma:flavor + flavor:acidity + aroma:acidity, family = binomial(link = "logit"),
    data = newdataset)

Deviance Residuals:
    Min       1Q    Median       3Q       Max
-3.6032  -0.1769    0.0000   0.2996    4.2438

Coefficients:
                                                  Estimate Std. Error z value
(Intercept)                                      -1.185e+03  3.116e+02  -3.802
country_of_originBurundi                          3.297e+00  9.721e+00   0.339
country_of_originChina                            4.982e-01  1.079e+00   0.462
country_of_originColombia                         1.779e+00  5.819e-01   3.057
country_of_originCosta Rica                       9.846e-02  7.742e-01   0.127
country_of_originCote d?Ivoire                   -1.037e+01  6.523e+03  -0.002
country_of_originEcuador                         -1.507e+00  1.507e+00  -1.000
country_of_originEl Salvador                      5.435e-01  9.684e-01   0.561
country_of_originEthiopia                         1.457e+01  6.167e+02   0.024
country_of_originGuatemala                       -5.908e-01  5.565e-01  -1.061
country_of_originHaiti                            2.546e+00  2.021e+00   1.260
country_of_originHonduras                        -3.395e-01  6.967e-01  -0.487
country_of_originIndia                           -2.671e+00  1.028e+00  -2.598
country_of_originIndonesia                       -1.766e-01  1.005e+00  -0.176
country_of_originKenya                            7.890e-01  1.583e+00   0.498
country_of_originLaos                            -1.456e+01  4.430e+03  -0.003
country_of_originMalawi                          -8.767e-01  1.308e+00  -0.670
country_of_originMauritius                       -9.658e+00  6.523e+03  -0.001
country_of_originMexico                          -7.121e-01  5.147e-01  -1.383
country_of_originMyanmar                         -1.701e+01  2.242e+03  -0.008
country_of_originNicaragua                        8.465e-01  2.158e+00   0.392
country_of_originPanama                           4.297e+00  2.199e+00   1.954
country_of_originPeru                            -1.417e+01  6.523e+03  -0.002
country_of_originPhilippines                      3.977e+00  4.846e+00   0.821
country_of_originTaiwan                           8.445e-01  6.920e-01   1.220
country_of_originTanzania, United Republic Of     1.122e+00  7.978e-01   1.407
country_of_originThailand                         2.870e+00  1.094e+00   2.623
country_of_originUganda                          -1.560e+00  7.724e-01  -2.020
country_of_originUnited States                   -5.293e-01  1.806e+00  -0.293
country_of_originUnited States (Hawaii)           1.460e+01  6.523e+03   0.002
```

```
country_of_originUnited States (Puerto Rico)  -2.897e+00  1.616e+00  -1.792
country_of_originVietnam                        2.389e+00  1.236e+00   1.933
country_of_originZambia                        -1.375e+01  6.523e+03  -0.002
aroma                                           1.197e+02  3.641e+01   3.287
flavor                                          1.221e+02  4.610e+01   2.648
acidity                                         5.099e+01  3.985e+01   1.280
category_two_defects                            5.617e-02  3.697e-02   1.519
altitude_mean_meters                            6.581e-04  3.289e-04   2.001
harvested2011                                  -1.242e-01  1.117e+00  -0.111
harvested2012                                   1.157e-01  9.627e-01   0.120
harvested2013                                   6.936e-01  9.752e-01   0.711
harvested2014                                   1.151e-01  9.805e-01   0.117
harvested2015                                  -4.723e-02  9.725e-01  -0.049
harvested2016                                   8.013e-01  1.020e+00   0.786
harvested2017                                   5.716e-01  1.028e+00   0.556
harvested2018                                   2.208e+00  1.284e+00   1.720
aroma:flavor                                   -1.199e+01  4.406e+00  -2.721
flavor:acidity                                 -2.941e+00  4.809e+00  -0.611
aroma:acidity                                  -3.083e+00  3.940e+00  -0.782
                                               Pr(>|z|)

(Intercept)                                    0.000144 ***
country_of_originBurundi                       0.734467
country_of_originChina                         0.644310
country_of_originColombia                      0.002235 **
country_of_originCosta Rica                    0.898811
country_of_originCote d?Ivoire                 0.998732
country_of_originEcuador                       0.317486
country_of_originEl Salvador                   0.574646
country_of_originEthiopia                      0.981155
country_of_originGuatemala                     0.288479
country_of_originHaiti                         0.207836
country_of_originHonduras                      0.626062
country_of_originIndia                         0.009367 **
country_of_originIndonesia                     0.860540
country_of_originKenya                         0.618170
country_of_originLaos                          0.997377
country_of_originMalawi                        0.502643
country_of_originMauritius                     0.998819
country_of_originMexico                        0.166547
country_of_originMyanmar                       0.993946
country_of_originNicaragua                     0.694802
country_of_originPanama                        0.050687 .
country_of_originPeru                          0.998266
country_of_originPhilippines                   0.411821
country_of_originTaiwan                        0.222329
country_of_originTanzania, United Republic Of  0.159470
country_of_originThailand                      0.008710 **
country_of_originUganda                        0.043389 *
country_of_originUnited States                 0.769400
country_of_originUnited States (Hawaii)        0.998213
country_of_originUnited States (Puerto Rico)   0.073120 .
country_of_originVietnam                       0.053296 .
country_of_originZambia                        0.998318
aroma                                          0.001014 **
```

```
flavor                                    0.008104 **
acidity                                   0.200709
category_two_defects                      0.128660
altitude_mean_meters                      0.045425 *
harvested2011                             0.911452
harvested2012                             0.904355
harvested2013                             0.476943
harvested2014                             0.906581
harvested2015                             0.961268
harvested2016                             0.432106
harvested2017                             0.578065
harvested2018                             0.085370 .
aroma:flavor                              0.006502 **
flavor:acidity                            0.540886
aroma:acidity                             0.433933
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1289.15  on 929  degrees of freedom
Residual deviance:  441.86  on 881  degrees of freedom
AIC: 539.86

Number of Fisher Scoring iterations: 17
```

```
plot_model(mod.cafe, show.values = TRUE,
           title = "", show.p = FALSE, value.offset = 0.5)
```
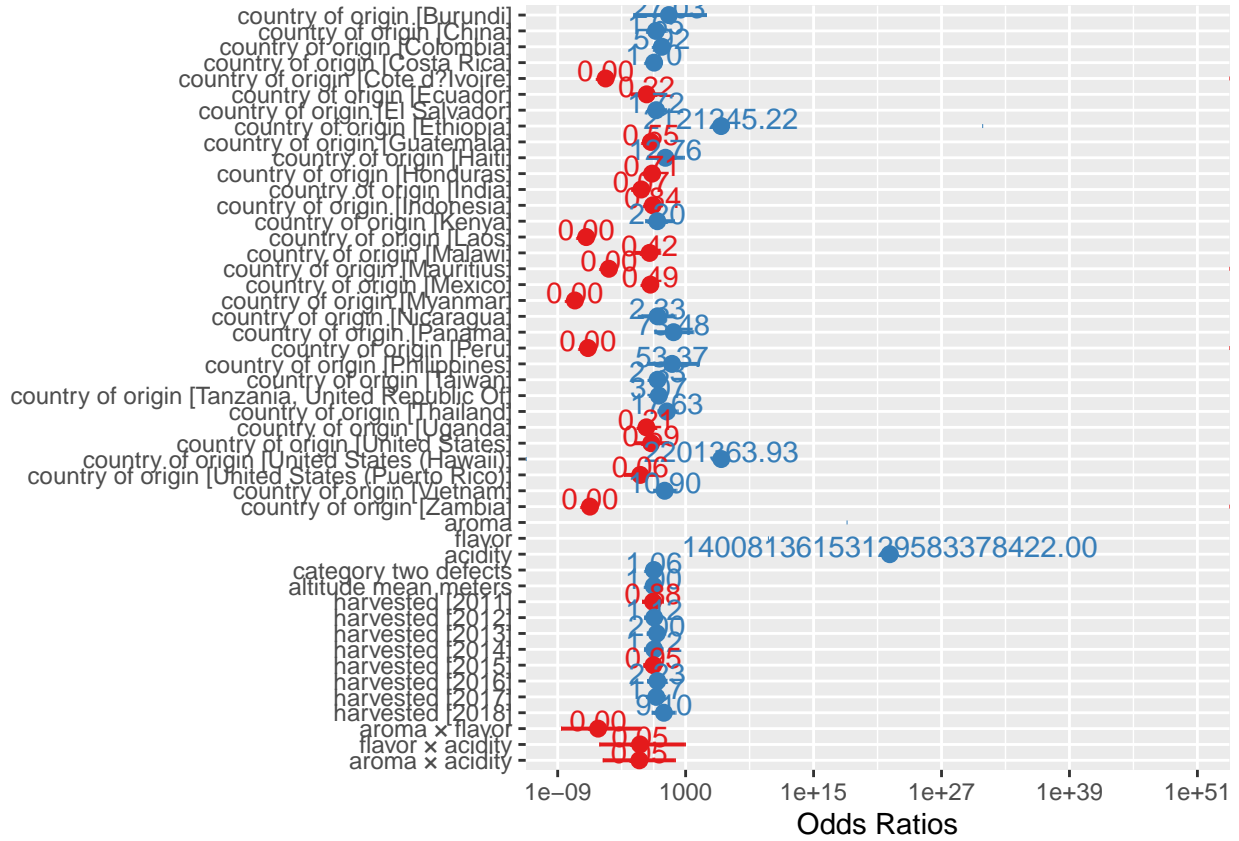

Figure 3: Odds of various factors influencing the quality of coffee(model with interaction terms)

In the results we can see the coefficients of aroma, flavor and acidity themselves are significantly positive, and there is a huge increase over the original model,while the coefficients of all our possible interaction terms are negative,which shows that these three variables may strongly moderate each other, indicating that their interaction actually affected the model at first.

After adding interaction terms, we can find that the AIC of the model decreases compared to the basic model, thus we can assume that the addition of the interaction terms improved our model.

GLM Stepwise After Adding Interaction Terms

In order to further improve our model, we fitted the GLM with interaction terms using the method of stepwise regression again with AIC as the criterion.

```
model.step <- stepAIC(glm(Qualityclass ~  country_of_origin +aroma + flavor+acidity+category_two_defects
summary(mod.cafe)#AIC532
```

```
Call:
glm(formula = Qualityclass ~ country_of_origin + aroma + flavor +
    acidity + category_two_defects + altitude_mean_meters + harvested +
    aroma:flavor + flavor:acidity + aroma:acidity, family = binomial(link = "logit"),
    data = newdataset)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.6032  -0.1769   0.0000   0.2996   4.2438

Coefficients:
                                                Estimate Std. Error z value
(Intercept)                                    -1.185e+03  3.116e+02  -3.802
country_of_originBurundi                        3.297e+00  9.721e+00   0.339
country_of_originChina                          4.982e-01  1.079e+00   0.462
country_of_originColombia                       1.779e+00  5.819e-01   3.057
country_of_originCosta Rica                     9.846e-02  7.742e-01   0.127
country_of_originCote d?Ivoire                 -1.037e+01  6.523e+03  -0.002
country_of_originEcuador                       -1.507e+00  1.507e+00  -1.000
country_of_originEl Salvador                    5.435e-01  9.684e-01   0.561
country_of_originEthiopia                       1.457e+01  6.167e+02   0.024
country_of_originGuatemala                     -5.908e-01  5.565e-01  -1.061
country_of_originHaiti                          2.546e+00  2.021e+00   1.260
country_of_originHonduras                      -3.395e-01  6.967e-01  -0.487
country_of_originIndia                         -2.671e+00  1.028e+00  -2.598
country_of_originIndonesia                     -1.766e-01  1.005e+00  -0.176
country_of_originKenya                          7.890e-01  1.583e+00   0.498
country_of_originLaos                          -1.456e+01  4.430e+03  -0.003
country_of_originMalawi                        -8.767e-01  1.308e+00  -0.670
country_of_originMauritius                     -9.658e+00  6.523e+03  -0.001
country_of_originMexico                        -7.121e-01  5.147e-01  -1.383
country_of_originMyanmar                       -1.701e+01  2.242e+03  -0.008
country_of_originNicaragua                      8.465e-01  2.158e+00   0.392
country_of_originPanama                         4.297e+00  2.199e+00   1.954
country_of_originPeru                          -1.417e+01  6.523e+03  -0.002
country_of_originPhilippines                    3.977e+00  4.846e+00   0.821
country_of_originTaiwan                         8.445e-01  6.920e-01   1.220
country_of_originTanzania, United Republic Of   1.122e+00  7.978e-01   1.407
country_of_originThailand                       2.870e+00  1.094e+00   2.623
country_of_originUganda                        -1.560e+00  7.724e-01  -2.020
country_of_originUnited States                 -5.293e-01  1.806e+00  -0.293
country_of_originUnited States (Hawaii)         1.460e+01  6.523e+03   0.002
country_of_originUnited States (Puerto Rico)   -2.897e+00  1.616e+00  -1.792
country_of_originVietnam                        2.389e+00  1.236e+00   1.933
country_of_originZambia                        -1.375e+01  6.523e+03  -0.002
```

```
aroma                                            1.197e+02  3.641e+01   3.287
flavor                                           1.221e+02  4.610e+01   2.648
acidity                                          5.099e+01  3.985e+01   1.280
category_two_defects                             5.617e-02  3.697e-02   1.519
altitude_mean_meters                             6.581e-04  3.289e-04   2.001
harvested2011                                   -1.242e-01  1.117e+00  -0.111
harvested2012                                    1.157e-01  9.627e-01   0.120
harvested2013                                    6.936e-01  9.752e-01   0.711
harvested2014                                    1.151e-01  9.805e-01   0.117
harvested2015                                   -4.723e-02  9.725e-01  -0.049
harvested2016                                    8.013e-01  1.020e+00   0.786
harvested2017                                    5.716e-01  1.028e+00   0.556
harvested2018                                    2.208e+00  1.284e+00   1.720
aroma:flavor                                    -1.199e+01  4.406e+00  -2.721
flavor:acidity                                  -2.941e+00  4.809e+00  -0.611
aroma:acidity                                   -3.083e+00  3.940e+00  -0.782
                                                 Pr(>|z|)
(Intercept)                                      0.000144 ***
country_of_originBurundi                         0.734467
country_of_originChina                           0.644310
country_of_originColombia                        0.002235 **
country_of_originCosta Rica                      0.898811
country_of_originCote d?Ivoire                   0.998732
country_of_originEcuador                         0.317486
country_of_originEl Salvador                     0.574646
country_of_originEthiopia                        0.981155
country_of_originGuatemala                       0.288479
country_of_originHaiti                           0.207836
country_of_originHonduras                        0.626062
country_of_originIndia                           0.009367 **
country_of_originIndonesia                       0.860540
country_of_originKenya                           0.618170
country_of_originLaos                            0.997377
country_of_originMalawi                          0.502643
country_of_originMauritius                       0.998819
country_of_originMexico                          0.166547
country_of_originMyanmar                         0.993946
country_of_originNicaragua                       0.694802
country_of_originPanama                          0.050687 .
country_of_originPeru                            0.998266
country_of_originPhilippines                     0.411821
country_of_originTaiwan                          0.222329
country_of_originTanzania, United Republic Of    0.159470
country_of_originThailand                        0.008710 **
country_of_originUganda                          0.043389 *
country_of_originUnited States                   0.769400
country_of_originUnited States (Hawaii)          0.998213
country_of_originUnited States (Puerto Rico)     0.073120 .
country_of_originVietnam                         0.053296 .
country_of_originZambia                          0.998318
aroma                                            0.001014 **
flavor                                           0.008104 **
acidity                                          0.200709
category_two_defects                             0.128660
```

```
altitude_mean_meters                          0.045425 *
harvested2011                                 0.911452
harvested2012                                 0.904355
harvested2013                                 0.476943
harvested2014                                 0.906581
harvested2015                                 0.961268
harvested2016                                 0.432106
harvested2017                                 0.578065
harvested2018                                 0.085370 .
aroma:flavor                                  0.006502 **
flavor:acidity                                0.540886
aroma:acidity                                 0.433933
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


(Dispersion parameter for binomial family taken to be 1)


    Null deviance: 1289.15  on 929  degrees of freedom
Residual deviance:  441.86  on 881  degrees of freedom
AIC: 539.86


Number of Fisher Scoring iterations: 17
```

```
plot_model(model.step, show.values = TRUE,
           title = "", show.p = FALSE, value.offset = 0.50)
```
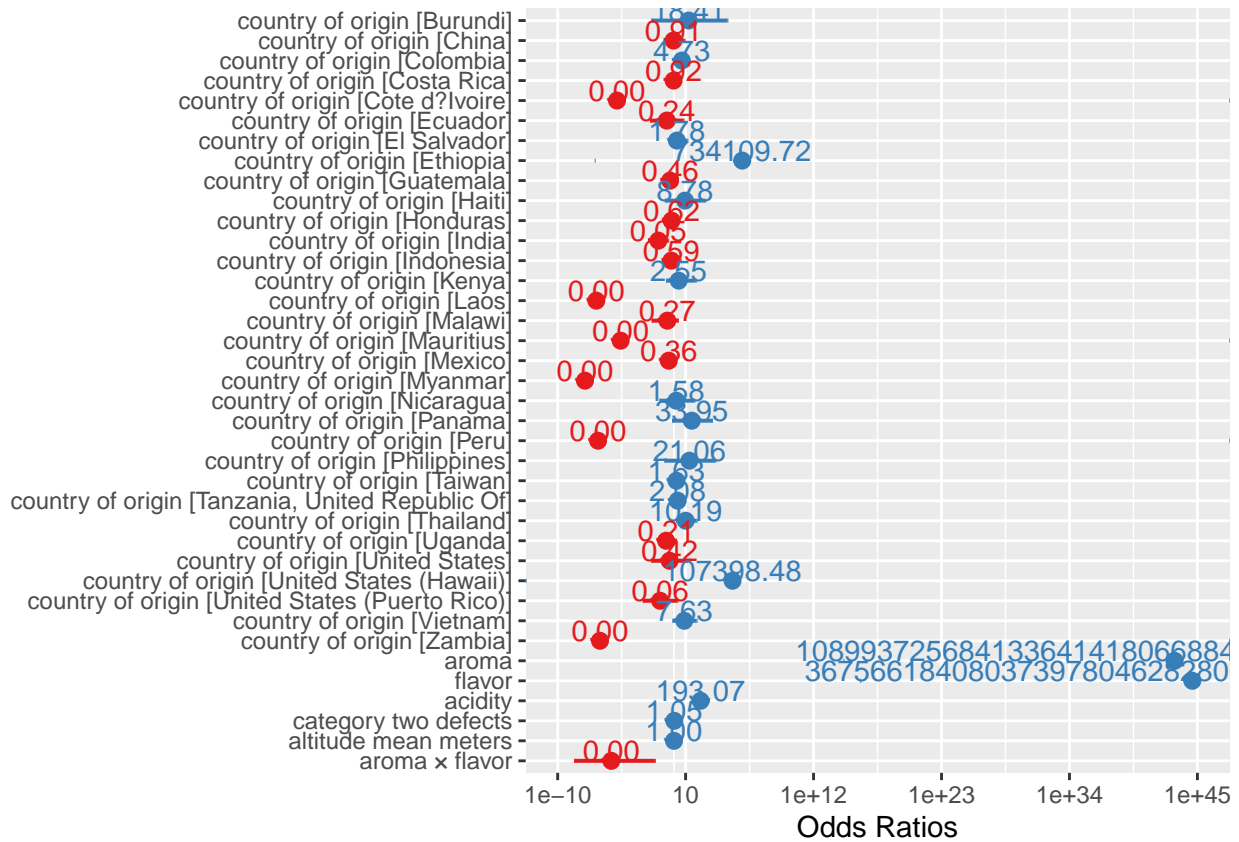

Figure 4: Odds of various factors influencing the quality of coffee(stepwise regression with interaction terms)

We can see from the results that the AIC decreased to the lowest among these four models we fitted. As AIC balances simplicity and accuracy when evaluating models, we can say that after adding an interaction term and doing the stepwise regression, our fourth model is the best model. Also, the last model has the lowest BIC=720, while the other three are 766,721,776 separately, which further demonstrates the superiority of our final model.

```
levels(newdataset$country_of_origin)
```

checking assumptions

Residuals Plots for each variables

```r
res <- resid(mod.cafe)
par(mfrow=c(3,2))
plot(newdataset$aroma,res,xlab='aroma')
abline(0,0)
plot(newdataset$flavor,res,xlab='flavor')
abline(0,0)
plot(newdataset$acidity,res,xlab='acidity')
abline(0,0)
plot(newdataset$category_two_defects,res,xlab='category two defects')
abline(0,0)
plot(newdataset$altitude_mean_meters,res,xlab='altitude mean meters')
abline(0,0)
plot(newdataset$harvested,res,xlab='harvested',ylab='res')
abline(0,0)
```
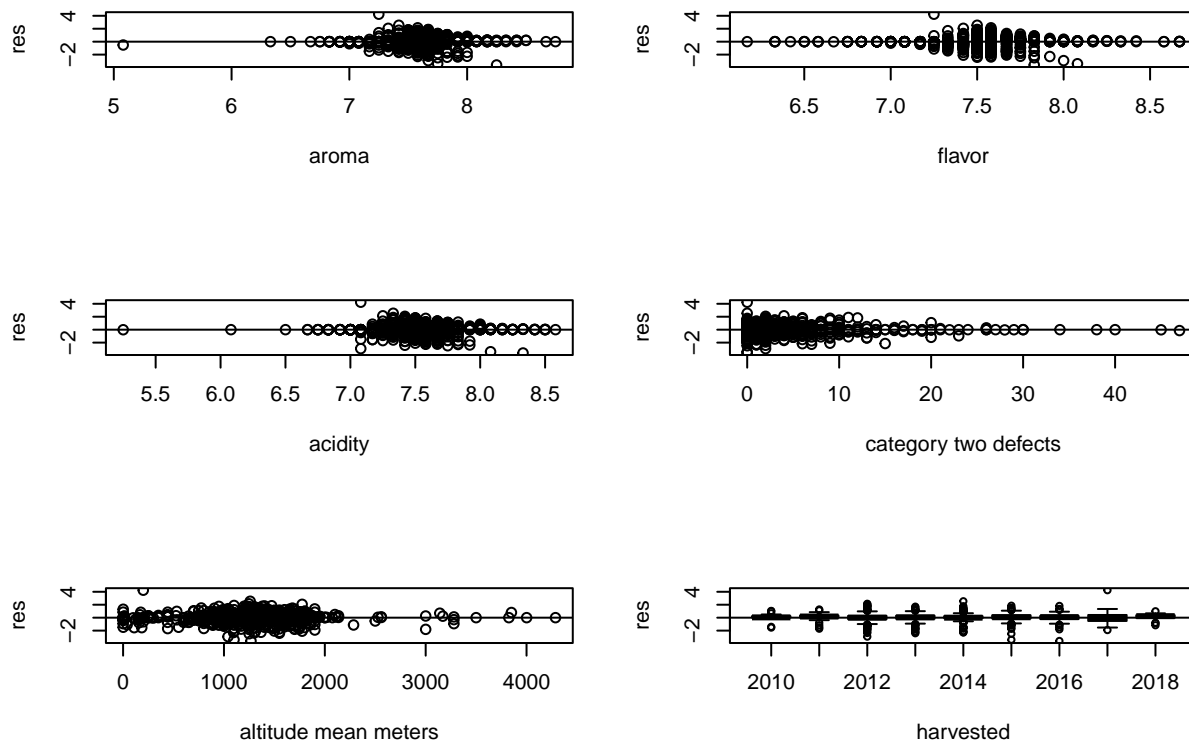

Figure 5: residuals against each variables

We see that there is an even spread of the residuals above and below the zero line for each variables,although there are a very few outlier points, overrall their spread on the graphs are acceptable, hence our assumption that the residuals have mean zero appears valid.

Density Plot

```
plot(density(res),xlab='residuals',title='')
```
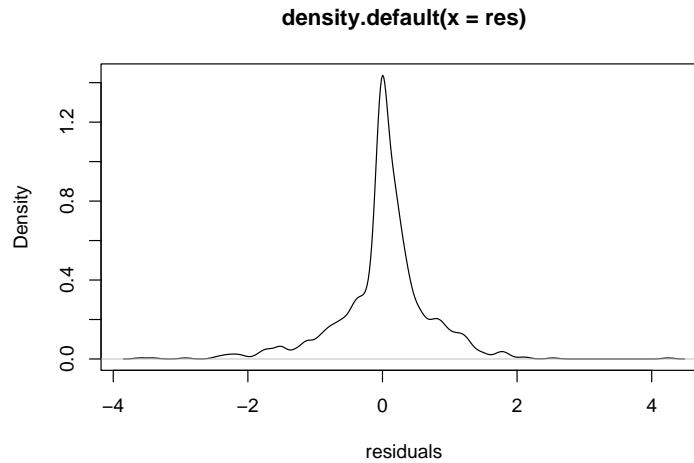
density.default(x = res)

Figure 6: density plot of residuals

In the graph we can see that the residuals are normally distributed with the mean 0, therefore the assumption is valid.

The remaining assumptions hold naturally at the time of our modelling, thus our model appears valid.

Conclusion

After data cleaning and processing of non-numerical data, we fitted the data to a regression model to observe the effect of each variable in the dataset on coffee quality, and we continued to improve the model by stepwise regression and adding possible interaction terms, resulting in the model with the smallest AIC value and therefore the most profile accurate. Looking at the summaries and graphs our final model, we can pick out the factors that have the greatest impact: aroma and flavor are very positively influencing on the quality of coffee, with coefficients of 99.1 and 102.62, acidity also appears to have relatively strong postive impact. The influence of origin varies very much, for example, Ethiopia and Hawaii contributes pretty much, for their coefficients are both more than ten(13.51 and 11.58). While Cote d'Ivoire,Laos,Mauritius, Myanmar,Peru and Zambia has less than -10 coefficients,making relatively strong negative effects. However, the many remaining origins do not seem to have much impact on the quality of the coffee.